1. Genetic variation (10pts)

The human genome is big (3 billion nucleotide, 2 copies) and genetic variations are arising constantly. These variations come in diverse forms:
(A) (2 pt) Name and define at least five major classes of human genome variation.

(B) (2 pt) For each class of variation, give an example of a mutational mechanism that could generate a variation of that class.

2. (2 pt) Several evolutionary and genetic factors shape diversity among human populations. Please name four such forces and briefly explain how these forces can affect genetic diversity.
3. (1 pt) Which of the following regarding somatic mutations/cells is NOT true?

(A) In humans, there is ~1 mutation per somatic cell division.
(B) Environmental exposures can cause different mutational signatures seen in normal cells across tissues.
(C) Different daughter cells from early somatic cell divisions can contribute unequally.
(D) Somatic mutations rates in humans are tissue-dependent, with the lowest rate seen in the spleen.
(E) 30X coverage is insufficient for the robust detection of somatic mutations from matched tumor-normal sequencing.

4. (1 pt) Which of the following statements is NOT true?

(A) Single nucleotide variants are the most common class of genetic variation and are very useful for genetic mapping.
(B) The current consensus germline mutation rate is ~1.2x10^{-8}.
(C) Retrogenes originate via the reverse transcription of mature messenger RNAs and usually result in genes without introns.
(D) Paternal age is positively associated with the germline mutation rate.
(E) Minisatellites mutate by replication slippage.

5. (2 pt) Human genome structural variation can be detected using four general methods: array comparative genomic hybridization, read-depth analysis of Illumina genome sequencing data, paired-end mapping analysis of genome sequencing data (a.k.a., read-pair analysis), and split-read mapping analysis of genome sequencing data.

(A) (0.4pt) Give a 1-2 sentence description of how each method is able to detect a structural variant.
(B) (0.4pt) Which of these four methods provides the highest genomic resolution?

(C) (0.4 pt) Which method(s) can detect balanced rearrangements such as inversions?

(D) (0.4 pt) Which method(s) can detect copy number variants (CNVs) such as deletions and duplications?

(E) (0.4 pt) Which method(s) can estimate the absolute copy number of a structurally variable genomic segment?
2. Population genetics (10pts)

1. (2 pt) You are rotating in a human genetic lab and your mentor asked you to check if a bi-allelic SNP violates the Hardy-Weinberg Equilibrium. The observed counts of the genotypes are provided in the table below. Please fill in the table, provide your calculation of the chi-squared test, and briefly explain your finding. (Hint: 3.84 is the critical value for \( \alpha = 0.05 \))

<table>
<thead>
<tr>
<th>WASHU-rs5488 Genotypes</th>
<th>T/T</th>
<th>T/C</th>
<th>C/C</th>
<th>HWE Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>500</td>
<td>50</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>Expected</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. (2 pts) 1 in 1800 US Caucasian newborns have cystic fibrosis. C is the normal allele and c is the risk allele. Individuals must be homozygous for the risk allele to have the disease.

(A) (1 pt) What percent of the above population have cystic fibrosis?

(B) (1 pt) Assuming a Hardy-Weinberg Equilibrium, how many newborns would have cystic fibrosis in a population of 10,000 people?
3. (6 pt) You got a call from Pierre Von Nostrumhauser, a distinguished but crazy French population geneticist who has been studying an isolated group of militant hillbillies. P.V. has discovered that these anti-social yokels are genetically identical except for a single biallelic locus (DLIVRANC) that is responsible for the ability to play the banjo. Strangely, this locus is not under selective pressure. P.V., a devout banjo enthusiast, is worried that the population may lose its musical abilities, so he asks you to investigate the decay of heterozygosity. The total population size is 200 individuals, and seems to have been this size for all of time.

(A) (2 pts) P.V. has sequenced the loci in 50 samples from the Georgian population. Calculate the heterozygosity, \( H \), of the population given the following initial distribution of alleles. (Hint: \( H_t = 1 - G_t \), where \( G_t \) is the homozygosity at generation \( t \))

<table>
<thead>
<tr>
<th>Allele</th>
<th>Number of alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-B (banjo skills)</td>
<td>11</td>
</tr>
<tr>
<td>D-NB (no banjo skills)</td>
<td>89</td>
</tr>
</tbody>
</table>

(B) (2 pts) Assuming a population size of 200, what is the expected heterozygosity, \( H \), in 6 generations? (Hint: \( H_t = (1-1/2N)^tH_0 \))

(iii) (2 pts) Determine the half-life of the banjo locus, i.e. the number of generations it will take for half of the heterozygosity to decay. Remember that \( t \approx 2N * \ln(2) \) where \( \ln(2) = 0.693 \). Discuss why P.V. should or should not be worried about the future of banjo playing.
3. Metagenomics (10pts)

1. You have performed 16S rRNA gene sequencing on fifteen mouse stool samples, with five each from: (A) untreated control mice (blue), (B) mice given 3 days of broad-spectrum antibiotics (red), and (C) mice given 3 days of antibiotics plus fecal microbiota transplant or FMT (green).

A) (1 pt) What kind of microbes are you sampling by 16S rRNA gene sequencing, and what microbes are you missing?

B) (1 pt) You perform a Shannon diversity analysis of your 16S rRNA sequencing results. What kind of a diversity metric is this?

C) (1 pt) What do the results of your Shannon diversity analysis, shown in the graph on the right, suggest about the biological effects of antibiotics treatment and fecal microbiota transplant (FMT)?

D) (1 pt) You next calculate UniFrac distances between your samples, and generate the following ordination plot. Based on this plot, which samples are most closely related, and which are most unrelated? What does this indicate about the communities present?
E) (1 pt) Why might an investigator choose to use metagenomic shotgun sequencing instead of 16S rRNA gene sequencing?

2. You are interested in identifying novel mechanisms of bacterial resistance to high salt concentrations. You have DNA isolated from a mixed community of “halophile”, or “salt-loving”, bacteria that thrive in high salt conditions, and you have identified in vitro salt conditions that halophiles tolerate but standard laboratory E. coli do not.

A) (2 pts) What would be an efficient way of identifying novel bacterial gene candidates from these halophiles that contribute to their tolerance of high salt conditions (one that does not rely on culturing new microbes)?

B) (1 pt) Once you have found halophile genes that contribute to tolerance of high salt conditions, how could you assess whether these were novel?

MULTIPLE CHOICE (only 1 possible correct answer):

3. (1 pt) Horizontal gene transfer between bacteria could occur by:
   A) Bacteriophage infection (transduction) of bacterial cells
   B) Uptake of small proteins by naturally transformable bacteria
   C) Fungal infection of the recipient bacteria
   D) Inheritance of a mutation after cell replication

4. (1 pt) Antibiotic resistance genes:
   A) All have very similar sequences, such that metagenomic shotgun sequencing is the only feasible way to identify them.
   B) Are absent in individuals from remote communities who have not been exposed to antibiotics.
   C) Have only been identified in human and animal microbiome samples.
   D) Tend to increase in patients given frequent courses of antibiotics.
4. Functional assays (10pts)

1. (3 pts) In the lectures we covered an approach to assay function called Sort-seq. Sort-seq uses flow cytometry to sort cells into bins based on levels of a fluorescent reporter, followed by sequencing to determine the distribution of library members across bins. Sort-seq is a popular method for conducting massively parallel functional assays. We discussed examples in which sort-seq is used for MPRAs, pooled CRISPR screens, and deep mutational scans, to measure the functional effects of enhancers, sgRNAs, and protein variants. **List two limitations of Sort-seq as method to measure function and briefly explain why they are limitations.**

2. (2 pts) In the assigned paper describing an MPRA to assess regulatory elements during neural induction (Inoue et al., 2019), the authors used lentivirus to integrate their MPRA library into the genomes of neural progenitor cells. List **one advantage and one limitation** of using lentiviral integration as a method to transduce MPRA libraries into cells. (You can contrast this method with plasmid-based MPRAs.)
3. (3 pts) Your collaborators have asked you to help design a Perturb-seq pooled CRISPR screen. They have designed a set of sgRNAs and they have expertise in single cell RNA-seq, but they need your assistance to implement the technology needed for assaying a pooled CRISPR library. They have asked you for advice on how design their CRISPR library construct so that the sgRNAs are expressed in each cell can be read out from single-cell sequencing data. Fortunately, you recently read the Perturb-seq paper (Dixit, et al. 2016) assigned in Genomics class. Briefly describe how Dixit and colleagues were able to identify the expressed sgRNAs from their single-cell sequencing data.

4. (2 pts) Deep mutational scans are a promising way to scale up testing of the impact of variants on protein function, and the resulting functional data can be used to assess whether a particular variant is pathogenic or benign. However, there are important limitations to the data from functional assays. Consider VAMP-seq, the assay used in the assigned paper Matreyek, et al. What was the protein feature/function assayed by VAMP-seq? List one limitation of using this feature as your readout of function in a deep mutational scan.
5. Variant analysis (10pts)

1. (4 pts) One important feature of the modeling approaches we discussed is that, in order to improve the prediction of variant effect, the model attempts to account for the assay itself. **Explain one reason why it is useful to explicitly include assay effects in your model.**

2. (3 pts) fitCons attempts to predict variant effects from what three kinds of input data? (List a minimum of two.)

3. (3 pts) We discussed one modeling approach (Faure, et al 2022) in which the authors used machine learning to construct a biophysical model that predicts variant effects on a protein-protein interaction domain. The biophysical model is based on two $\Delta G$'s: one for folding and one for binding. These authors argue that modeling biophysical effects that underlie a phenotype will improve predictions of variant effects because (choose one):

   A) You cannot understand the effect of a variant on disease unless you that variant’s detailed molecular properties.

   B) Different underlying biophysical effects often lead to the same phenotype. Resolving this ambiguity will improve predictions of the combined effects of multiple variants in the same gene.

   C) Biophysical predictions are more accurate than ‘black box’ machine learning predictions because we understand the underlying mechanism.
6. Comparative genomics (10pts)

1. (3 points)
In Statistical Coupling Analysis, the covariance across aligned columns (positions) is used to generate a large matrix, which is then used to identify non-contiguous ‘sectors’ (after decomposition). Answer the questions about the Multi-Sequence Alignment below.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>Axolotl</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>Goose</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>Hatchimal</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>Reef Shark</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>Ringed Octopus</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>Chinchilla</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>C</td>
<td>Capped conure</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>Cone Snail</td>
</tr>
</tbody>
</table>

- List all the columns (positions) that are invariant (conserved):
- List the pair of columns that are most-strongly coupled:

2. (2 points)
Comparative Genomics of Primates and Humans. Circle all the answers below that are related to the concept of “Selective Sweep” abbreviated below as SS.

1 SS indicates strong immune response
2 SS is usually caused by transposons
3 SS is related to high Ka/Ks (non-synonymous / synonymous)
4 SS is Related to low Ka/Ks (non-synonymous / synonymous)
5 SS is usually caused by neuronally-related genes
6 The gene FoxP2 has mutations and/or regions which are an example of SS
7 SS is evidence of advantageous alleles in a region
8 SS is evidence of embryonic lethal mutations
9 SS only occurs in Homo sapiens genomes
3. (5 points)
While trying to generate a Hidden Markov Model (HMM) to represent a new type of CRISPR gRNA Spacer and PAM site you explore the models below. Listed first are 8 example genomic target sequences. (you want the HMM to identify the sequences listed below)

- The numbers listed in the boxes below (next to the A/C/G/T) are called __? (Circle one). (Cumulative probabilities, Transition probabilities, Transversion abilities, Emission probabilities, State ranks)

- Which of the HMMs listed below would be the most effective at identifying the flavor of sequences (spacer/PAM) listed above? B